Bayesian Break-Point Forecasting in Parallel Time Series, with Application to University Admissions

D. B. Rubin; T. W. F. Stroud

The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 15, No. 1. (Mar., 1987), pp. 1-19.

The Canadian Journal of Statistics
Vol. 15, No. 1, 1987, Pages 1–19
La Revue Canadienne de Statistique

Bayesian break-point forecasting in parallel time series, with application to university admissions

D.B. RUBIN and T.W.F. STROUD*

Harvard University and *Queen's University*

Key words and phrases: Short history forecasting, pooled cross-sectional time series, predictive means and variances, change-point models.
AMS 1985 subject classifications: Primary 62M20; secondary 62F15.

ABSTRACT

A regular supply of applicants to Queen's University in Kingston, Ontario is provided by 65 high schools. Each high school can be characterized by a series of grading standards which change from year to year. To aid admissions decisions, it is desirable to forecast the current year's grading standards for all 65 high schools using grading standards estimated from past year's data. We develop and apply a Bayesian break-point time-series model that generates forecasts which involve smoothing across time for each school and smoothing across schools. "Break point" refers to a point in time which divides the past into the "old past" and the "recent past" where the yearly observations in the recent past are exchangeable with the observations in the year to be forecast. We show that this model works fairly well when applied to 11 years of Queen's University data. The model can be applied to other data sets with the parallel time-series structure and short history, and can be extended in several ways to more complicated structures.

RÉSUMÉ

Une bonne partie de la clientèle de l'université Queen's (à Kingston, en Ontario) provient chaque année des mêmes écoles secondaires. Le niveau moyen de préparation des finissants de ces 65 écoles peut être représenté numériquement par un indice dont la valeur change d'année en année. Pour faciliter l'évaluation des dossiers, on désirait prévoir cette valeur pour chacune des écoles à partir des données des années antérieures. A cette fin, nous avons élaboré et testé un modèle bayésien de ces séries chronologiques d'indices. Les données ont été lissées afin de tenir compte des variations dans le temps et entre les écoles. Des points de rupture permettant de discerner entre le passé récent et ancien ont également été incorporés au modèle et les données des années récentes ont été considérées comme échangeables avec les observations prédites. Ce modèle donne des résultats acceptables pour les données accumulées pendant 11 ans par l'université Queen's. Il se prête bien aux situations où l'on a affaire à plusieurs petites séries chronologiques parallèles et il peut être généralisé de diverses manières afin de tenir compte de structures plus compliquées.

1. INTRODUCTION

1.1. The Problem.

For a number of years, Queen's University in Kingston, Ontario has been interested in monitoring the relative grading standards of those high schools which provide a regular

*Supported by the Natural Sciences and Engineering Research Council, Grant A7220.

supply of applicants. The concern is that the grades of some high schools may be inflated relative to those of others. Since Queen's University does not use a standardized admissions test such as the SAT (Scholastic Aptitude Test), which is required by many U.S. universities, decisions are based primarily on high-school grades, and consequently differences in the relative grading standards of high schools tend to give unfair advantages to students from high schools with lower grading standards.

In previous work (Rubin and Stroud 1977), a method was introduced for calculating *relative grading standards* from the high-school final year average (HSA) of numerical marks and the first-year Queen's average (FYA) of students in a specified set of high schools over a sequence of matriculation years. The actual technique, described in the next section, is based on a modelling of FYA given HSA, and generates a value for each school-year combination in the time span under consideration. These high-school standards cannot be calculated for any *current* applicant, since none of these applicants has an FYA score. Our concern here is with extrapolating the past standards of each high school one more year, into the year of current applicants. This forecasting work is important because, if the forecasts are successful, they provide the university admissions office with supplementary information helpful in interpreting the HSAs provided by the different high schools. In some cases, this information might be used in deciding whom to admit and whom not to admit.

The collection of relative grading standards, evaluated for each high school over the same sequence of matriculation years, comprises what we call a set of parallel time series. In this article, we develop a Bayesian model for such data with an unknown break point, which incorporates smoothing across parallel time series. Based on this model, described in Section 1.3, the forecasted relative grading standard for each school is a weighted average of the school's standards in past years, where more recent years always receive more weight than more distant years. Such a weighting is intuitively quite attractive, and the fact that it results from a formal Bayesian model means that standard errors can be attached to the forecasts and that model deficiencies can be assessed. The appeal of our method, which provides forecasts using decreasing data-based weights over time, is somewhat similar to the appeal of the use of empirical Bayesian methods in variance-component models, which provide a way of using the data to obtain smoothed estimates.

1.2. Relative Grading Standards.

The relative grading standards, shown in Table 1, were calculated according to the method described in Rubin and Stroud (1977), which involves fitting a linear model each matriculation year relating the first-year Queen's University average of numerical marks (FYA) to the final-year high-school average (HSA), the square of the final-year average, and dummy variables representing the various schools in the set. The parameter estimates representing the schools are then centered so that their weighted average is zero, the results being called the relative grading standards of these schools. The difference between the standards of two schools in a given matriculation year equals the difference between the estimated expected first-year university averages of students from the two schools having the same final year high school average.

We see from Table 1 that there is a great deal of fluctuation in each school's standard over time. For some schools, however, the values are consistently negative, while for others they are positive. This finding is important for forecasting.

Other statistics relevant to the production of Table 1, namely estimated regression coefficients, residual mean squares, and numbers of students per school, are described in Rubin and Stroud (1984, Appendix C).

1.3. The Bayesian Break-Point Model.

Our method of forecasting current relative grading standards is based on a Bayesian "break-point" model, which specifies one distribution for the "old" past and another distribution for the "recent" past, where the break point between old and recent must be estimated. More specifically, let X_{st} be the relative grading standard for school s at time t, $s = 1, \ldots, S$, $t = 1, \ldots, T$, where the objective is to forecast the current relative grading standards $X_{s,T+1}$ for $s = 1, \ldots, S$. The X_{st} beginning with the unknown break point are assumed to follow the familiar linear model underlying a one-way ANOVA (analysis of variance), where the groups are the high schools and the grading standards within a group are independently and identically distributed replications: for $t \geq k$, $X_{st} \sim \mathbf{N}(\mu_s, \sigma^2)$. The data before the break point also follow the same model, but with different parameter values: for $t < k$, $X_{st} \sim \mathbf{N}(\mu_s^*, \sigma^{*2})$. The set of grading standards to be forecast $\{X_{s,T+1}; s = 1, \ldots, S\}$ are being modelled as the set consisting of one future observation in each group with the same parameter values $(\mu_1, \ldots, \mu_S; \sigma^2)$ as the observed data from the break point onward. This establishes the simple break-point time-series aspects of our model, which generates estimates of current standards that are weighted averages of past standards with more weight given to the standards from the recent past.

Our model also has Bayesian aspects which smooth the estimates across the parallel time series. The μ_s are related by a prior distribution, and the μ_s^* are also related by a prior distribution; specifically, $\mu_s \sim N(\mu_0, \tau^2)$ and $\mu_s^* \sim N(\mu_0^*, \tau^{*2})$. In the analysis of the Queen's data, $\mu_0 = \mu_0^* = 0$, but this is not necessary in general. Furthermore, we make our analysis fully Bayesian by specifying prior distributions for all parameters k, σ^2, σ^{*2}, τ^2, τ^{*2}. Details of these specifications and subsequent derivations of related posterior distributions are presented in Section 3.

1.4. Comments on the Bayesian Model.

If only one high school were involved, with no obvious trend in time, we would consider the average of the "recent" past to be a satisfactory predictor. When the series is quite stable in time, the recent past should include the entire history and a reasonable predictor would be the average value over this past; if there were an abrupt change three years ago, the recent past would include only the past three years of data. Thus, one of the major objectives of our analysis is to quantify evidence in the data about the duration of the recent past. In so doing, we decide how to trade off the increased bias of prediction but decreased variance that arises from going back too far into the past against the reduced bias but increased variance that arises from basing predictions on too few observations.

Since the Queen's data set involves parallel time series at many schools, we consider it advantageous to shrink our predictors toward a common value. The purpose of such shrinking is to make use of the information in the entire data set to produce estimates at particular schools that are less extreme than the original estimates, which were based on only a few observations. Sometimes called "borrowing strength", such shrinkage is the essential feature of James-Stein estimation, empirical Bayes estimation, and hierarchical Bayesian inference, and has become a dominant theme in recent statistical literature; see e.g. Mosteller and Wallace (1964), Jackson, Novick, and Thayer (1971), Lindley and Smith (1972), Efron and Morris (1975), Rubin (1980), Morris (1983).

1.5. Outline.

Section 2 presents results of analyses of the Queen's data. These results indicate that our model works reasonably well. For example, the average actual squared forecast error

TABLE 1: Relative grading standards of 65 high schools.

School	Year 1	2	3	4	5	6	7	8	9	10	11
1	−1.700	−1.889	−5.106	−2.140	−2.497	−1.247	0.603	−4.114	−3.419	−1.258	−1.660
2	1.311	−2.660	−3.325	−1.437	0.114	1.853	−1.841	−2.257	2.274	−1.879	0.360
3	−3.383	−5.192	1.735	7.852	−2.651	5.553	−5.611	4.333	−0.705	−1.483	−1.957
4	−0.196	−1.853	−6.624	−7.941	−4.630	−8.579	−3.502	−8.578	−6.893	−13.381	−7.110
5	2.407	1.741	1.604	−0.518	0.255	−1.800	1.427	3.454	−0.090	−0.638	0.304
6	3.517	4.001	0.017	0.084	1.842	0.639	4.954	3.582	1.589	4.597	4.728
7	5.561	−2.227	4.735	1.881	6.644	3.933	−4.311	−0.344	3.061	5.868	4.164
8	−1.974	−5.276	−6.455	0.941	−2.738	−3.171	3.501	3.545	1.657	6.464	3.417
9	0.294	1.826	−0.151	2.752	−3.112	0.484	3.300	4.208	−1.290	1.837	4.137
10	2.446	−0.247	3.467	2.163	−0.161	−7.759	−1.711	0.758	0.719	−5.659	−0.719
11	−0.981	3.225	−1.595	1.737	3.497	−4.738	1.234	−2.265	−0.131	0.517	−1.404
12	−2.847	0.038	−1.947	−0.047	1.197	−3.654	−1.733	−1.151	−1.347	−2.459	−4.492
13	0.927	4.738	2.395	−2.331	7.220	2.731	5.060	1.476	−0.176	0.412	−4.938
14	−1.730	3.668	−0.887	4.471	6.187	3.488	2.382	1.749	4.997	0.528	5.122
15	−7.005	−3.005	−0.512	1.981	0.502	1.336	4.042	−0.185	−0.402	5.325	4.742
16	0.966	−3.183	2.042	1.272	−0.653	−1.953	5.992	3.435	−2.650	1.808	−2.048
17	−2.467	4.223	2.589	1.636	3.310	4.206	−2.477	3.030	5.189	1.851	4.397
18	0.000	0.592	3.075	0.309	0.009	0.801	−1.476	1.485	0.450	2.630	−1.077
19	1.039	−3.025	−2.742	1.478	−2.873	−5.739	−2.557	−0.063	−3.939	−6.096	−6.749
20	1.610	−2.904	−1.220	−0.469	1.246	−0.404	−0.252	−11.627	0.553	−3.127	1.126
21	−2.783	−2.784	2.049	1.653	−3.498	0.488	−3.525	−6.876	−5.622	−1.269	−3.910
22	−0.899	0.241	0.343	0.224	0.783	0.544	2.157	5.218	−6.676	5.119	3.787
23	−3.396	−1.922	−2.273	−1.965	−5.433	−6.333	1.990	−3.696	5.334	−13.158	−4.822
24	−3.002	2.544	0.935	3.301	1.120	1.738	2.326	1.781	−1.917	−1.557	−1.352
25	−1.934	−1.821	3.064	−6.000	2.932	0.679	0.945	−5.315	0.745	−2.480	−2.595
26	−0.736	0.072	−1.180	−3.582	0.629	−2.763	−5.299	1.213	−6.874	2.643	−4.865
27	0.000	2.349	1.468	0.696	−2.620	−0.396	−0.630	2.815	0.825	0.598	−0.823
28	3.539	−3.260	2.580	3.016	0.198	3.793	5.276	1.852	2.806	0.788	4.902
29	−5.152	−3.247	2.282	−4.188	−1.974	2.832	−3.718	−7.988	−3.838	2.636	−3.221
30	1.670	5.085	5.146	0.292	3.413	−1.150	−6.250	0.804	4.089	0.547	4.482
31	5.381	1.622	5.716	1.770	3.789	10.869	7.325	8.391	10.437	16.360	4.165

32	−4.596	6.825	3.166	1.123	−0.583	5.725	2.432	−0.884	4.635	−2.067	−6.736
33	−7.069	−14.985	0.374	−5.307	0.694	−2.022	−3.425	−2.427	−0.988	−0.939	−1.693
34	3.382	0.191	−3.432	−10.424	−7.302	−12.309	−6.569	−0.378	2.979	11.403	3.898
35	−0.699	−0.449	2.485	1.958	7.424	4.683	4.692	4.676	4.065	5.803	6.375
36	5.789	2.362	1.148	3.446	5.963	1.881	6.954	1.261	5.242	1.224	−2.692
37	−15.134	−1.842	−3.922	3.926	−0.294	−6.669	4.241	2.442	−0.134	0.119	6.083
38	6.200	1.122	0.394	0.932	−2.065	1.366	−5.803	−6.431	−9.648	2.706	−0.910
39	−8.257	0.889	3.558	1.308	2.798	0.563	0.681	−1.472	6.108	1.251	−20.936
40	3.633	1.367	−6.843	−11.409	−10.802	−0.857	6.486	−2.607	1.597	−3.900	−0.169
41	2.438	−4.705	−0.741	−0.348	2.506	4.626	1.112	1.636	1.517	1.964	4.077
42	−0.474	0.071	5.221	1.886	−1.840	−1.194	−7.988	2.502	−1.578	1.961	−3.117
43	−5.759	0.098	1.318	5.465	0.553	−7.497	−1.859	−3.695	3.384	−0.830	−4.046
44	7.469	2.395	9.974	6.957	0.497	3.181	5.809	5.387	4.391	0.084	3.644
45	1.047	−3.227	−0.043	3.399	−1.695	0.926	−1.637	11.947	1.047	3.007	1.942
46	−5.681	4.416	−1.966	−3.891	−1.753	7.181	−3.066	−2.409	−6.015	2.367	0.968
47	3.541	4.152	0.290	−2.478	−2.786	−2.106	−8.647	−13.817	−1.400	−11.772	−0.945
48	0.927	−0.205	3.509	−1.156	4.256	−1.904	6.540	−1.195	−0.773	2.719	−0.473
49	1.524	4.514	−0.878	7.431	1.387	−4.374	−5.601	2.556	4.607	10.438	−11.362
50	2.976	−2.548	5.084	3.036	4.640	−1.059	6.169	−5.538	8.105	13.477	5.647
51	−0.250	1.276	−0.647	−1.438	6.784	−0.261	−1.865	−0.266	−3.279	−3.031	−0.657
52	0.569	10.522	2.801	−6.450	−0.186	1.096	0.768	−2.601	−0.201	−6.806	−0.231
53	1.513	5.134	−3.314	4.207	−6.010	−4.603	−2.416	−1.687	−4.201	−7.320	0.456
54	−4.841	0.506	2.395	4.263	0.362	8.520	2.439	−3.614	3.340	1.499	−4.131
55	2.784	−1.250	−7.708	−1.254	6.782	−0.171	4.783	12.389	3.307	5.670	0.772
56	−1.853	−0.427	3.507	5.546	2.706	2.446	−0.056	2.538	0.282	−1.635	−0.350
57	−1.430	−1.533	−6.849	4.032	−3.145	−6.333	−9.819	3.801	3.076	5.212	−1.330
58	1.190	7.258	11.985	−2.531	−3.281	−13.220	4.308	−3.868	−2.867	3.942	0.853
59	−0.417	1.715	6.162	−0.335	−0.801	−5.894	−3.814	−4.828	4.052	10.019	−2.572
60	0.224	0.707	−3.596	−6.860	0.103	2.002	2.791	7.681	5.940	2.219	6.524
61	5.840	5.616	−4.076	1.341	0.908	0.260	−0.615	−3.351	−7.538	1.348	0.057
62	−4.198	−0.804	−3.518	6.805	5.473	1.696	−4.645	0.733	−8.835	−10.756	1.671
63	1.368	−11.765	0.087	−2.349	−7.040	−3.450	−10.235	−9.734	1.387	−0.935	−4.860
64	−11.791	−0.295	0.416	−3.590	4.572	3.741	−0.234	−4.189	1.177	1.166	−1.998
65	3.843	−3.573	−3.049	−3.196	1.201	−0.929	−2.196	2.127	−2.608	−3.294	3.127

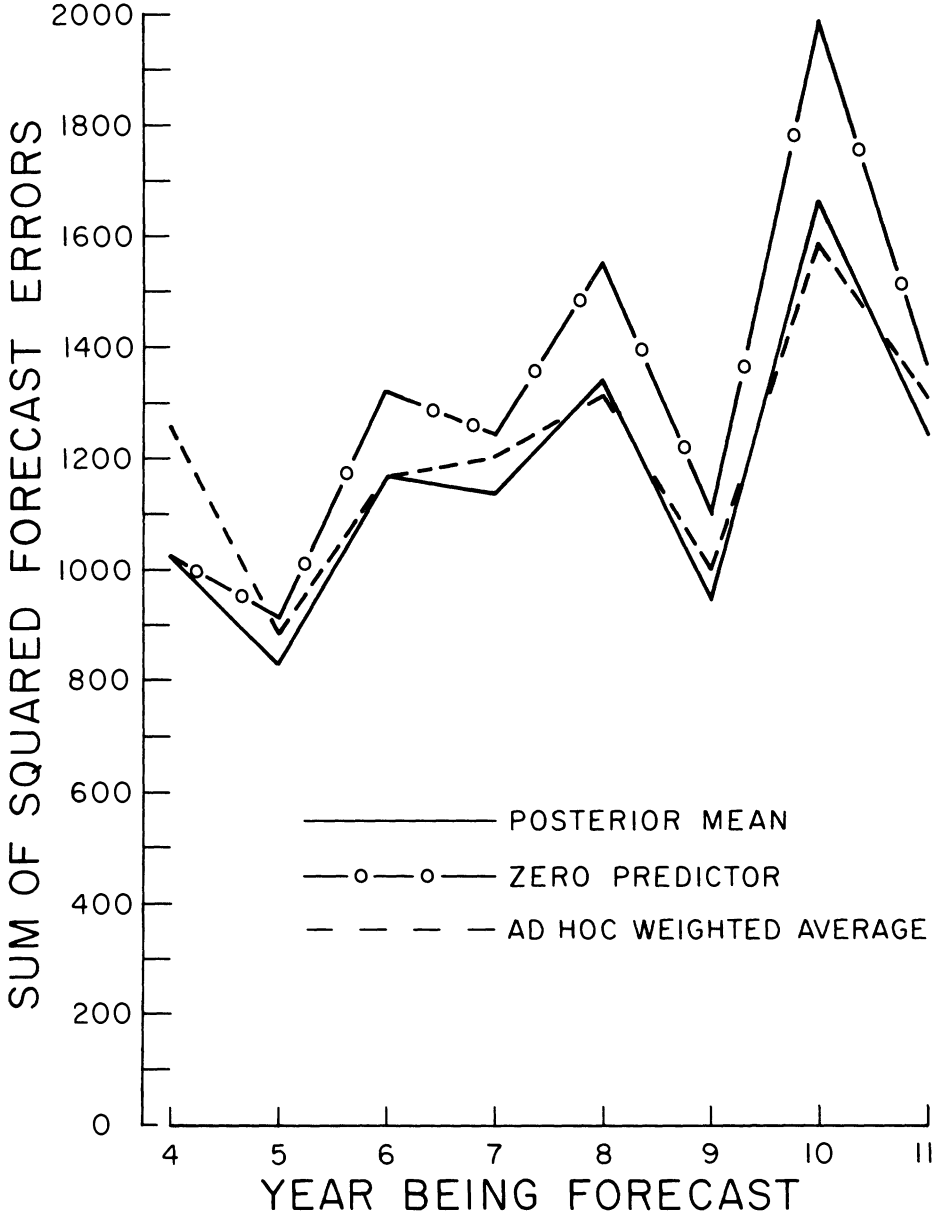

Figure 1: Comparison of sums of squared forecast errors for three techniques used for forecasting relative grading standards.

across all schools and years, based on the data from Table 2 shown in Figure 1, is 11% less for our technique (posterior means) than for the grand mean (zero) (which is what would be used if past data were ignored).

Section 3 presents the model more formally and includes derivations of relevant results. The presentation in Section 4 addresses the propriety of the model's assumptions as applied to the Queen's data. Finally, Section 5 discusses generalizations and other approaches to this problem.

TABLE 2: Sum of squared forecast errors for forecasting by straight averaging and by the posterior mean, the zero predictor, and the linearly increasing (*ad hoc*) weighted average.

	Straight average												
Year being predicted	Beginning with year 1	2	3	4	5	6	7	8	9	10	Pos. Mean	Zero Pred.	Wted. Ave.
4	1326	1330	1470	0	0	0	0	0	0	0	1023	1025	1257
5	941	912	983	1323	0	0	0	0	0	0	827	913	883
6	1250	1237	1217	1252	1310	0	0	0	0	0	1167	1321	1167
7	1177	1204	1292	1362	1371	1885	0	0	0	0	1137	1245	1202
8	1351	1378	1368	1267	1414	1542	1855	0	0	0	1340	1547	1311
9	982	993	988	1073	1108	1191	1301	1934	0	0	947	1099	997
10	1638	1651	1606	1655	1645	1666	1637	1563	1861	0	1656	1986	1580
11	1274	1340	1345	1326	1313	1321	1400	1577	1928	2539	1245	1371	1309

2. FORECASTS OF RELATIVE GRADING STANDARDS

2.1. Posterior Predictive Means.

A Bayesian model can be used to forecast future values, such as grading standards, using the posterior predictive distribution of these values. In our model, the posterior predictive mean of $X_{s,T+1}$ given μ_s is simply μ_s. Thus, the posterior predictive mean of $X_{s,T+1}$ equals the mean of the posterior distribution of μ_s; this is derived in Section 3 and Appendix A. Using the data in Table 1, we can create forecasts using our model for year $T + 1$ from data for years $1, \ldots, T$. To demonstrate the performance of our method, such forecasts were made for a range of values of T, corresponding to the use of the technique over the years as data accumulated, beginning with a very short history ($T = 3$) and ending with the forecasting of the eleventh year ($T = 10$). The year-4 forecasts were obtained from the first three years' data, the year-5 forecasts were obtained from the first four years' data, etc. In the next two subsections, the performance of these forecasts is discussed.

2.2. Evaluating the Posterior Mean Forecasts.

Table 2 displays the sum of squared forecast errors (actual minus predicted) over the 65 schools. The first ten columns of the body of the table give sums of squared errors of point predictors defined as averages of the recent past for various fixed starting points in time; thus, column 1 shows forecasts based on the entire past, column 2 shows forecasts based on the past beginning with year 2, etc. The last three columns show the sum of squared forecast errors by three forecast techniques: the posterior mean (the method described herein), a forecast of zero grading standard for each school, and an *ad hoc* weighted average technique where the past years $1, 2, \ldots, T$ are simply given the weights $1, 2, \ldots, T$, respectively. (Figure 1 shows the same information.) In any appropriate application of the method of this article, the past data should provide information about how the past should be weighted, so that our method should perform better in the long run than some general rule where the weights do not depend on the data. Table 2 shows that the posterior mean forecast was better than a grand mean (zero) forecast in all eight forecasting trials; hence the grading standards do not behave like white noise. Notice, however, that a forecast based on only the last year of the data set performed worse than the zero forecast six times out of eight. In comparing the posterior mean with the *ad hoc* weighting scheme, the posterior mean did better on five occasions and worse on two occasions, with one tie. Table 3 compares the posterior mean, the zero forecast, and the *ad hoc* weighting scheme using the sum of absolute errors instead of sum of squared errors. The results are similar.

2.3. Posterior Predictive Intervals.

For each year being forecasted (4 through 11), 95% prediction intervals were computed according to the formula

$$\text{predictive mean} \pm 2(\text{predictive variance})^{\frac{1}{2}},$$

where the predictive mean equals the posterior mean and the predictive variance is given by (A.6). Table 4 gives, for each year, the number of actual values that fell below the lower prediction limit and the number of actual values that fell above the upper prediction limit. These results indicate that these nominal 95% intervals may be safely interpreted for these data as 90−95% intervals, a performance which we consider reasonably satisfactory.

TABLE 3: Sum of absolute errors for forecasting by the posterior mean, the zero predictor, and the weighted average.

	Sum of errors		
Year being predicted	Post. Mean	Zero Pred.	Wted. Ave.
4	194.9	200.2	214.5
5	184.0	186.9	192.5
6	200.0	220.4	200.5
7	220.7	238.0	216.6
8	225.6	244.3	224.5
9	196.6	211.0	196.8
10	238.6	257.8	239.4
11	197.6	218.9	197.6

TABLE 4: Performance of 95% prediction intervals for the next year's relative grading standards of 65 schools.

Year being predicted	Actual below lower limit	Actual between limits	Actual above upper limit
4	3	60	2
5	1	64	0
6	4	58	3
7	4	60	1
8	2	60	3
9	2	63	0
10	4	56	5
11	2	63	0

It is also useful to note that schools whose grading standards are substantially above or below the average (zero) can be identified from the forecasts, which are listed in Rubin and Stroud (1984, Table 2). For example, in year 11 there are eight schools whose predicted standards exceed 2.5 in magnitude. (A difference of 2.5 in relative grading standards refers to a difference of 2.5 in expected first-year average on a numerical scale from 0 to 100.) In all these eight cases, the sign of the actual relative grading standard (Table 1) corresponds with the sign of the forecast. In this way, the admissions office can identify those schools whose grading standards are the lowest and those whose grading standards are the highest.

3. FORMAL STATEMENT OF MODEL

3.1. Initial Specifications.

Let X_{st} be the value at school s and time t, where we have data for $s = 1, \ldots, S$ and $t = 1, \ldots, T$. Our objective is to forecast the S values of X_{st} for $t = T + 1$.

Let k be the first value of t in the "recent" past, so that $t = 1, \ldots, k - 1$ indexes the old past and $t = k, \ldots, T$ indexes the recent past. We mean formally by this statement that given the vector parameter $\theta = (k, \mu_1^*, \ldots, \mu_S^*, \mu_1, \ldots, \mu_S, \sigma^{*2}, \sigma^2)$, the observations X_{st} for $s = 1, \ldots, S$ and $t = 1, \ldots, T$ are independent with

$$X_{st} \sim N(\mu_s^*, \sigma^{*2}), \qquad s = 1, \ldots, S, \quad t = 1, \ldots, k-1, \tag{3.1}$$

and

$$X_{st} \sim N(\mu_s, \sigma^2), \qquad s = 1, \ldots, S, \quad t = k, \ldots, T, T+1. \tag{3.2}$$

3.2. Adding Bayesian Structure.

Thus far, the only ties across the S time series are via the common variances σ^{*2} and σ^2 and the common break point k. With the model (3.1)−(3.2) and k known, an obvious predictor of $X_{s,T+1}$ is

$$\frac{1}{T-k+1} \sum_{t=k}^{T} X_{st}.$$

Because there are many parallel series, however, in order to improve the estimates of each $X_{s,T+1}$ by borrowing strength across the series, we use a hierarchical Bayesian model with prior distributions on the $\mu_s^*$and the μ_s. The quantities μ_s for $s = 1, \ldots, S$ are taken to be independently and identically distributed with mean μ_0 and variance τ^2; similarly, the $\mu_s^*$are independently and identically distributed $\mathbf{N}(\mu_0^*, \tau^{*2})$. Here μ_0 and μ_0^* are known (zero in our application), and an asterisk continues to denote a before-break quantity. In order to complete a fully Bayesian specification, we need to place a distribution over τ^{*2}, τ^2, σ^{*2}, σ^2, and k.

3.3. The Prior Distribution of the Variances Given the Break Point.

For convenience, we replace the hyperparameters σ^2 and τ^2 by λ and ϕ, where $\lambda = 1/\sigma^2$ and $\phi = (1 + n\tau^2/\sigma^2)^{-1}$, with $n = T - k + 1$. Similarly, we replace σ^{*2} and τ^{*2} by λ^* and ϕ^*, where $\lambda^* = 1/\sigma^{*2}$ and $\phi^* = (1 + n^*\tau^{*2}/\sigma^{*2})^{-1}$, with $n^* = k - 1$. Given k, the quantities λ, λ^*, ϕ, and ϕ^* are *a priori* independent with

$$p(\lambda, \lambda^*, \phi, \phi^*) = p_B(\lambda^*, \phi^*)p_A(\lambda, \phi), \tag{3.3}$$

where

$$p_A(\lambda, \phi) = (1 - r)\phi^{-r}\left(\frac{\nu_0 m_0}{2}\right)^{\nu_0/2} \lambda^{(\nu_0/2)-1} e^{-\lambda\nu_0 m_0/2}\{\Gamma(\nu_0/2)\}^{-1} \tag{3.4}$$

and

$$p_B(\lambda^*, \phi^*) = (1 - r)\phi^{*-r}\left(\frac{\nu_0 m_0^*}{2}\right)^{\nu_0/2} \lambda^{*(\nu_0/2)-1} e^{-\lambda^*\nu_0 m_0^*/2}\{\Gamma(\nu_0/2)\}^{-1}, \tag{3.5}$$

$$0 < \phi \leq 1, \quad 0 < \phi^* \leq 1, \quad \lambda > 0, \quad \lambda^* > 0.$$

The factors p_B and p_A refer to before and after the break, respectively. If $k = 1$, there is no break and the factor p_B is unity. The priors on λ and λ^* are scaled chi-squared priors with ν_0 degrees of freedom, corresponding to scaled inverted chi-squared priors on σ^2 and σ^{*2}. The prior mean of the precision λ is $1/m_0$, corresponding to m_0 as a prior estimate of σ^2 (and similarly for λ^*, m_0^*, and σ^{*2}). The form of the transformations from τ^2 and τ^{*2} to ϕ and ϕ^*, respectively, and of the priors on ϕ and ϕ^*, are based on Strawderman (1971).

These priors are proper. As $\nu_0 \to 0$ and $r \to 1$, the limiting forms are improper and are equivalent to the priors used by Box and Tiao (1968; 1973, Sections 5.2, 7.2). We do not use the improper priors on λ and λ^*, because they do not yield proper posterior distributions when there is only one time period before or after the break. We do not use the improper priors on ϕ and ϕ^*, because a zero posterior probability on values of $k > 1$ would result when $r = 1$, because the factor $1 - r$ in p_A when $k = 1$ is much larger than the factor $(1 - r)^2$ in $p_A p_B$ when $k > 1$.

3.4. Specifying Hyperparameters in the Prior Distribution of the Variances.

Lacking specific prior information on the hyperparameters ν_0, m_0, m_0^*, and r, we suggest that they be chosen in the manner described in the following paragraphs. Most of these choices are dependent on the data.

Since the prior expectation of ϕ is $(1 - r)/(2 - r)$, we suggest estimating ϕ from the data and solving for r, viz. $r = (1 - 2\hat{\phi})/(1 - \hat{\phi})$. For the estimate $\hat{\phi}$, one may use the MLE based on the entire set of observed data and a variance-components model with known grand mean μ_0; thus

$$\hat{\phi} = \mathrm{MSW}/\mathrm{MSB}$$

where $\mathrm{MSW} = \Sigma\Sigma(X_{st} - \bar{X}_{s\cdot})^2$ and $\mathrm{MSB} = T\Sigma(\bar{X}_{s\cdot} - \mu_0)^2$. However, small values of r produce a prior structure that tends to shrink the μ_s, μ_s^*too much in the direction of μ_0, without paying sufficient attention to the data. For this reason we suggest a lower bound of $\frac{1}{2}$ for r. Thus take

$$r = \max\left\{\frac{1}{2}, \frac{1 - 2\hat{\phi}}{1 - \hat{\phi}}\right\}.$$

Note that r will be $>\frac{1}{2}$ provided $\hat{\phi} < \frac{1}{3}$. In the analysis of the Queen's University data, this method applied to the first ten years data yielded $r = 0.54418$, which was used in all eight forecasts.

Concerning the hyperparameters of the prior distributions of λ and λ^*, we suggest using $\nu_0 = 6$, because this is the smallest value (i.e., corresponding to the least informative prior) that results in finite posterior means and variances when $r \geq \frac{1}{2}$. For choosing m_0 and m_0^*, we suggest $m_0 = \Sigma_{s=1}^{S} (X_{sT} - \bar{X}_{\cdot T})^2/(S - 1)$, a between-group mean square based on only the final time period, and $m_0^* = \Sigma_{s=1}^{S} (X_{s1} - \bar{X}_{\cdot 1})^2/(S - 1)$ based on the $t = 1$ data. If the between-group variance is zero ($\phi = 1$), then m_0 is an unbiased estimate of σ^2; otherwise it will tend to overestimate σ^2. We use this conservative procedure because we wish to avoid underestimating variances. When different values of k are combined in the final analysis, the information in the data will dominate to yield a reasonable posterior distribution of σ^2.

3.5. Posterior Distributions Given k.

Given k, the posterior distributions of σ^2 and the μ_s and hence the predictive distributions of the s values of $X_{s,T+1}$ can be found from a standard Bayesian analysis which is very similar to the analysis presented in Box and Tiao (1968; 1973, Section 7.2). The data from this analysis, given k, constitute the $S \times n$ data matrix of the recent past, and the model is the simple one-way layout with S groups and n observations per group. The so-called "random effects" version of the model is the one used here, because of the normal prior on the μ_s.

3.6. The Prior Distribution of k.

As a prior distribution for k over the integers $1, 2, \ldots, T$ we suggest, unless circumstances indicate otherwise, the discrete uniform prior with a constant probability of $1/T$.

This was the prior used for the analysis of the Queen's University admission data of Section 2. Whatever prior is used for k, this prior, together with the results conditional on k which have been derived above, determine the posterior distribution of k, as described in Appendix A. In Appendix A it is also explained how to obtain the posterior means and variances of the μ_s and the predictive variances of the $X_{s,T+1}$.

4. DISCUSSION OF ASSUMPTIONS OF THE MODEL

4.1. Introduction.

As with any model applied to real data, there exist important questions concerning the propriety of the assumptions. Since our model has complicated structure, it is appropriate to discuss why we think the assumptions are reasonable for the Queen's University admissions data.

4.2. One Common Break Point.

First, we consider the assumption of a single break point k, common to all schools. The conditions that determine what we have called grading standards at a particular school tend to change from time to time, but there is no guarantee they will have changed only once during the period covered by the data. However, because the Bayesian prediction scheme using a single unknown break point k produces, before smoothing across schools, a weighted average of past data with increasing weights, and because the history is short, we feel that the predictive accuracy of the results would not be changed much by incorporating more break points, so we have chosen a single break point for simplicity. Likewise, we feel that allowing the break point to change from school to school would not improve predictions sufficiently to warrant the added complexity.

4.3. Independent Normal Distributions.

Regarding the model of independent normal observations before and after the break, the normality can be regarded as an expression that the average of the recent past should be a reasonable predictor if we knew the break point. The assumption of independence across schools is an expression of the idea of exchangeability among schools; such exchangeability is usually implied by the notion of the S time series being "parallel".

4.4. Common Variances.

Regarding the common variance of σ^2 after the break and the common variance σ^{*2} before the break, which represent the variation from year to year of the standards of a given school, we note that this variation comes from two sources: actual changes in school standards from year to year due to things like variations of curricula, teaching methods, and examinations, and the variance of estimation of parameters in the linear model from which the school standards are obtained. Although the numbers of students in the various schools provide some information about the latter effects, the former effects could be determined only from the school standards themselves. Since a model incorporating both kinds of effects would be quite complicated, and since separate unknown variances for each school would require the estimation of many parameters, we feel that it is appropriate to use the common unknown within-group variance specified by the model even though its estimate will tend to be high for that small collection of schools with relatively large numbers of students.

4.5. *Prior Distributions for μ_s and μ_s^*.*

The normality with zero means of the prior distributions of the μ_s, μ_s^*seems appropriate here, since the standards for all schools appear symmetrically distributed about their central value, which is by definition close to zero; we note also that no schools show up as "outliers".

4.6. *Prior Distribution of ϕ.*

The predictions will depend somewhat on the prior distribution of ϕ (or, equivalently, of τ^2), because this prior distribution dictates how much shrinking towards the central value μ_0 will take place. We feel that a uniform prior on τ^2 [which was used by Leonard (1976), and shrinks approximately the same amount as the James-Stein estimator] does not create enough shrinkage for the purposes of the present problem. There is a great deal of random fluctuation in the data analyzed here, and predictions that are heavily based on individual school means with little shrinkage would not be expected to do as well overall as predictions with a fair amount of shrinkage toward zero. The family of priors which we use has a larger amount of shrinkage, and when $\frac{1}{2} \leq r < 1$, the prior produces minimax point estimates of means in the one-way model with known variances σ^2 (Strawderman, 1971) for $n \geq 5$.

4.7. *Prior distribution on* k.

We wish to express prior ignorance of where the break point is, if it exists ($k > 1$), and also allow for the possibility of no break at all ($k = 1$). We feel that equal prior probabilities of all values of k from 1 to T is a reasonable way of expressing this.

5. GENERALIZATIONS AND EXTENSIONS

5.1. *More General Parallel Time Series.*

The ideas of this article may be applied to other collections of "parallel time series" where it is desired to forecast the next value for each series in the collection. Examples of parallel time series are not difficult to find, e.g., burglary rates in 20 cities or motor-vehicle death rates in 48 states. This kind of data structure is common in the econometric literature, where it is known as "pooled cross-sectional and time-series data" (e.g. Griffiths and Anderson 1982). We expect our model to be most applicable when the number of series is between 20 and 100.

If the series contain obvious time trends, these trends may be estimated and subtracted from the series before the beginning of the analysis, as was done with the relative grading standards. Sometimes there may not be any obvious way to estimate the grand mean for the data set as a whole. We may want to include the grand mean as a hyperparameter, and possibly to allow its value to be different after the break from before the break. The theory for this formulation is discussed in Appendix B.

5.2. *Inclusion of Global Predictor Variables.*

Another possible extension is to the case involving a global external predictor variable, available at each point in time. Such an extension follows immediately, in theory at least, from the work of this paper together with the results presented in Stroud (1984). The predictor could be a useful global predictor, or it could be time itself, allowing each series to have its own slope, as well as its own mean.

5.3. More Than One Break Point.

Finally, we mentioned that some sets of parallel time series may cover a sufficiently long time span that we may prefer to use only a recent section of the data instead of the whole data set, to avoid estimating a break point near the beginning of the data set, when a more recent but less marked break point is more relevant to predicting the future. One way of determining an appropriate value of T is to make one or more trial runs, the first one using all available time periods, and note on each occasion the posterior probabilities, and particularly the posterior mode, of k. If, on the first run, this mode is far enough back in time that the choice of this $k = \hat{k}$ as the break point would jeopardize the detection of a more recent and more relevant break point, then the time periods $1, 2, \ldots, \hat{k} - 1$ should be deleted for the second run. If the resulting modal break point is still too far back in time, then this procedure can be repeated, as often as is necessary.

5.4. Related Methods.

A Bayesian treatment of a break point as an unknown parameter in a single time series was studied by Smith (1980). For parallel time series, Thisted and Wecker (1981) used shrinkage toward a central value but did not utilize a break point, using instead exponential smoothing to provide differential weights for past values. A more comprehensive version of a correlated-means model which yields differential weights using dynamic linear modelling is described by West, Harrison, and Migon (1985). So far, results are only available for the case of a single time series.

Swamy and Mehta (1975, 1977) have presented a random-coefficient regression model approach to parallel time series, which in principle could be applied to the Queen's University data, since HSA is being used as a predictor for FYA. This formulation, however, requires the estimation of a large number of parameters whose sampling variability is not fully represented in any implied predictive intervals.

APPENDIX A. CALCULATIONS OF POSTERIOR AND PREDICTIVE MEANS AND VARIANCES

In this section, the joint posterior distribution, given the break point k, of the μ_s ($s = 1, \ldots, S$), λ, and ϕ together with the μ_s^*, λ^*, and ϕ^* is presented, along with explicit formulae for the conditional posterior means and variances of the μ_s, given k, and expressions for the marginal density of the data set, with μ_s^*, μ_s, λ^*, λ, ϕ^*, and ϕ integrated out. These expressions determine the posterior distribution of k. Finally, formulae for the posterior means and variances of the μ_s based on the full posterior distributions (i.e., of all parameters including k) are derived, and from these the predictive mean and variance of the next observation in each time series are obtained.

The joint posterior density of $(\boldsymbol{\mu}, \lambda, \phi)$, given the after-break data y, is obtained from the model and prior specifications given in the previous section as

$$p(\boldsymbol{\mu}, \lambda, \phi | y) = p(\boldsymbol{\mu} | \lambda, \phi; y) p(\lambda, \phi | y). \tag{A.1}$$

The first factor of the right-hand side of (A.1) is a slight modification of Box and Tiao [1973, (7.2.12)], reflecting knowledge of the grand mean. It is given by

$$p(\boldsymbol{\mu} | \lambda, \phi, y) \sim \mathbf{N}((1 - \phi)\bar{y} + \phi\mu_0\mathbf{1}, (n\lambda)^{-1}(1 - \phi)I), \tag{A.2}$$

where $\bar{y}$ is the S-dimensional vector of after-break averages. The second factor is obtained as

$$p(\lambda, \phi | y) = \frac{p(\lambda, \phi) p(y | \lambda, \phi)}{p(y)}, \tag{A.3}$$

where $p(y) = \int\int p(\lambda, \phi) p(y | \lambda, \phi) \, d\lambda \, d\phi$. The second factor in the numerator of the right-hand side of (A.3) comes from

$$p(y | \lambda, \phi) = \int p(\boldsymbol{\mu} | \lambda, \phi) p(y | \boldsymbol{\mu}, \lambda, \phi) \, d\boldsymbol{\mu}.$$

Combining $p(y | \lambda, \phi)$ with the joint prior $p(\lambda, \phi)$ given by (3.4), we obtain

$$\begin{aligned} p(\lambda, \phi) p'y | \lambda, \mu_0, \phi) = {} & \frac{(1 - r)(\nu_0 m_0/2)^{\nu_0/2} \lambda^{(\nu_0/2) - 1} e^{-\lambda \nu_0 m_0/2}}{\phi^r \Gamma(\nu_0/2)} \\ & \times \frac{\lambda^{nS/2}}{(2\pi)^{nS/2} |I_{nS} + J_n \otimes n^{-1}(\phi^{-1} - 1) I_S|^{1/2}} \\ & \times \exp\left(\frac{\lambda}{2}(y - \mu_0 \mathbf{1}_{nS})' \{I_{nS} + J_n \otimes n^{-1}(\phi^{-1} - 1) I_S\}^{-1} (y - \mu_0 \mathbf{1}_{nS})\right) \end{aligned}$$

where I_n is the $n \times n$ identity and J_n is the $n \times n$ matrix of ones.

The integral of this expression, $p(y)$ [or more correctly $p(y|n)$ or $p(y|k)$] is provided below; it is needed to get the posterior distribution of k:

$$\begin{aligned} p(y|n) = {} & \frac{(1 - r)(\nu_0 m_0)^{\nu_0/2} \Gamma((\nu_2/2) + 1 - r) \Gamma(\{(\nu_0 + \nu_1)/2\} - (1 - r))}{\pi^{nS/2} \Gamma(\nu_0/2)(\nu_0 m_0 + \nu_1 m_1)^{(\nu_0 + \nu_1 - 2 + 2r)/2} (\nu_2 m_2)^{(\nu_2 + 2 - 2r)/2}} \\ & \times I_w\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1}{2} - (1 - r)\right), \end{aligned} \tag{A.4}$$

where $w = \nu_2 m_2/(\nu_0 m_0 + \nu_1 m_1 + \nu_2 m_2)$, $\nu_1 = (n - 1)S$, $\nu_2 = S$, $\nu_1 m_1 = \Sigma\Sigma(y_{st} - \bar{y}_{s\cdot})^2$, $\nu_2 m_2 = n\Sigma_s(\bar{y}_s - \mu_0)^2$, and I_w represents the incomplete beta function. The symbols y_{st} and $\bar{y}_{s\cdot}$ denote elements and row averages, respectively, of the after-break data matrix y.

The posterior mean and variance, given k, of the μ_s $(s = 1, \ldots, S)$ may be obtained in a manner similar to the last paragraph of Box and Tiao (1973, Section 7.2.3). We have

$$\mathscr{E}(\mu_s | y) = \bar{y}_{s\cdot} - \mathscr{E}(\phi | y)(\bar{y}_{s\cdot} - \mu_0), \tag{A.5}$$

and

$$\begin{aligned} Var(\mu_s | y) = {} & \frac{(\nu_0 m_0 + \nu_1 m_1)\{1 - \mathscr{E}(\phi|y)\} + \nu_2 m_2 \{\mathscr{E}(\phi|y) - \mathscr{E}(\phi^2|y)\}}{n(\nu_0 + \nu_1 + \nu_2 - 2)} \\ & + (\bar{y}_{s\cdot} - \mu_0)^2 Var(\phi|y), \end{aligned}$$

where

$$\mathscr{E}(\phi^q|y) = \frac{(1 - w)^q B\left(\frac{\nu_2}{2} + q + 1 - r, \frac{\nu_0 + \nu_1}{2} - (q + 1 - r)\right) I_w\left(\frac{\nu_2}{2} + q + 1 - r, \frac{\nu_0 + \nu_1}{2} - (q + 1 - r)\right)}{w^q B\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1}{2} - (1 - r)\right) I_w\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1}{2} - (1 - r)\right)}$$

$q = 1, 2$; and $Var(\phi|y) = \mathscr{E}(\phi^2|y) - [\mathscr{E}(\phi|y)]^2$. Note that, in all these expressions, conditioning on k is implied, since y is the after-break data matrix.

To obtain the posterior distribution of k, we first note that $p(x|k) = p(y^*|n^*)p(y|n)$, where $n^* = k - 1$, $n = T - k + 1$, y^* is the before-break data matrix, y is the after-break data matrix, and the full data matrix $\boldsymbol{x}$ has been partitioned into $\boldsymbol{y}^*$ and $\boldsymbol{y}$ according to the value of the break point k. Here $p(\boldsymbol{y}|n)$ is given by (A.4), and for $k > 1$, $p(\boldsymbol{y}^*|n^*)$ is given by (A.4) with asterisks on the quantities n, m_0, m_1, m_2, and ν_1, defined for the before-break data set the same way the unasterisked quantities were defined for the after-break data set. For the case $k = 1$ we define $p(\boldsymbol{y}^*|n^*) = 1$, since here $p(\boldsymbol{x}|k) = p(\boldsymbol{y}|n)$.

Now the posterior distribution of k, for $k = 1, 2, \ldots, T$, is obtained from

$$p(k|x_{11} \ldots x_{ST}) = \frac{p(k)p(x_{11}, \ldots, x_{ST}|k)}{\sum_{k=1}^{T} p(k)p(x_{11}, \ldots, x_{ST}|k)}.$$

The posterior mean and variance of the μ_s in the full analysis may now be determined. For example, we have

$$\mathscr{E}(\mu_s|x_{11} \ldots x_{ST}) = \sum_{k=1}^{T} p(k|x_{11}, \ldots, x_{ST})\mathscr{E}(\mu_s|\boldsymbol{y}_k)$$

where $\boldsymbol{y}_k = \boldsymbol{y}$ is the after-break data, conditioned on k, and $\mathscr{E}(\mu_s|\boldsymbol{y}_k)$ is given by (A.5). Similarly

$$\begin{aligned} Var(\mu_s|x_{11} \ldots x_{ST}) &= \mathscr{E}\{Var(\mu_s|\boldsymbol{y}_k)\} + Var\{\mathscr{E}(\mu_s|\boldsymbol{y}_k)\} \\ &= \sum_k Var(\mu_s|\boldsymbol{y}_k)p(k|x_{11} \ldots X_{ST}) + \sum_k \mathscr{E}^2(\mu_s|\boldsymbol{y}_k)p(k|x_{11} \ldots x_{ST}) \\ &\quad - \mathscr{E}^2(\mu_s|x_{11} \ldots x_{ST}). \end{aligned}$$

The predictive mean of a new observation $x_{s,T+1}$ is the same as the posterior mean of μ_s. The predictive variance can be shown to be

$$Var\{x_{s,T+1}|x_{11} \ldots x_{ST}\} = \mathscr{E}(\sigma^2|x_{11} \ldots x_{ST}) + Var\{\mu_s|x_{11} \ldots x_{ST}\}. \tag{A.6}$$

Thus all that is needed further is $\mathscr{E}(\sigma^2|x_{11}, \ldots, x_{ST}) = \mathscr{E}\{(1/\lambda)|x_{11}, \ldots, x_{ST}\}$. This also may be obtained by conditioning on k and on ϕ. Because, for the one-way ANOVA model with n observations, the posterior distribution of λ is a scaled chi-squared with $\nu_0 + \nu_1 + \nu_2$ degrees of freedom and scale factor

$$\frac{1}{\nu_0 m_0 + \nu_1 m_1 + \phi\nu_2 m_2},$$

it follows that

$$\mathscr{E}(\sigma^2|\boldsymbol{y}, k, \phi) = \frac{\nu_0 m_0 + \nu_1 m_1 + \phi\nu_2 m_2}{\nu_0 + \nu_1 + \nu_2 - 2},$$

so that, with $m_1 = m_1(\boldsymbol{y}_k)$ and $m_2 = m_2(\boldsymbol{y}_k)$,

$$\mathscr{E}(\sigma^2|x_{11}, \ldots, x_{ST}) = \sum_k p(k|x_{11}, \ldots, x_{ST}) \times \frac{\nu_0 m_0 + \nu_1 m_1(\boldsymbol{y}_k) + \nu_2 m_2(\boldsymbol{y}_k)\mathscr{E}(\phi|k)}{\nu_0 + \nu_1 + \nu_2 - 2}.$$

APPENDIX B. MODEL AND RESULTS WHEN THE PRIOR MEANS ARE NOT KNOWN IN ADVANCE

The development of Section 3 was motivated by the application to high-school grading standards, where there is a built-in centering about the value $\mu_0 = 0$. In many situations such a centering may not exist. Here the grand mean before the break and the grand mean after the break may well differ. In this section we give the modifications to the theory that follow if the grand mean μ_0^* before the break and grand mean μ_0 after the break are unknown instead of known.

The prior distribution of the μ_s and of the $\mu_s^*$are as specified in Section 3.2 except that the hyperparameters μ_0^*, μ_0, τ^{*2}, and τ^2 are now all unknown. In place of (3.3), (3.4), and (3.5) we have

$$p(\mu_0, \mu_0^*, \lambda, \lambda^*, \phi, \phi^*) = p_B(\mu_0^*, \lambda^*, \phi^*) p_A(\mu_0, \lambda, \phi), \tag{B.1}$$

where

$$p_A(\mu_0, \lambda, \phi) = \frac{(1 - r)\left(\dfrac{\nu_0 m_0}{2}\right)^{\nu_0/2} \lambda^{(\nu_0 - 1)/2} \exp\left\{-\dfrac{\lambda}{2}\left(\nu_0 m_0 + \dfrac{(\mu_0 - \xi)^2}{c}\right)\right\}}{(2\pi c)^{1/2} \phi^r \Gamma\left(\dfrac{\nu_0}{2}\right)} \tag{B.2}$$

and

$$p_B(\mu_0^*, \lambda^*, \phi^*) = \frac{(1 - r)\left\{\left(\dfrac{\nu_0 m_0^*}{2}\right\}^{\nu_0/2} \lambda^{*(\nu_0 - 1)/2} \exp\left\{\dfrac{\lambda^*}{2}\left(\nu_0 m_0^* + \dfrac{(\mu_0^* - \xi)^2}{c^*}\right)\right\}}{(2\pi c^*)^{1/2} \phi^{*r} \Gamma\left(\dfrac{\nu_0}{2}\right)}, \tag{B.3}$$

$0 < \phi \leq 1, \quad 0 < \phi^* \leq 1, \quad \lambda > 0, \quad \lambda^* > 0, \quad -\infty < \mu_0 < \infty, \quad -\infty < \mu_0^* < \infty.$

The derivation of the posterior distributions of Appendix A is carried through with the pair of hyperparameters (λ, ϕ) replaced by the triple (λ, ϕ, μ_0). These changes lead to the following expression replacing (A.4):

$$p(y|n) = \frac{(1 - r)(\nu_0 m_0)^{\nu_0/2} \Gamma((nS + \nu_0)/2)}{\pi^{nS/2} c^{1/2} \Gamma(\nu_0/2)} A(y; c, r, \xi, n, \nu_0, m_0, S), \tag{B.4}$$

where

$$A(y; c, r, \xi, n, \nu_0, m_0, S)$$

$$= \int_0^1 \frac{\phi^{s/2 - r}\, d\phi}{(nS\phi^{-1} + c)^{\frac{1}{2}}[\nu_0 m_0 + \nu_1 m_1 + \phi \nu_2 m_2 + (\bar{y} - \xi)^2\{c + (nS\phi)^{-1}\}^{-1}]^{(nS + \nu_0)/2}} \tag{B.5}$$

where $\bar{y}$ is the grand mean of the after-break data set, and ν_2, m_2 have been changed to $\nu_2 = S - 1$, $\nu_2 m_2 = n\Sigma_s(\bar{y}_s - \bar{y})^2$, respectively. If c is finite, a numerical integration is required to evaluate (B.5). As c gets large, the term c^{-1} in $nS\phi + c^{-1}$ becomes negligible and so does $(\bar{y} - \xi)^2\{c + (nS\phi)^{-1}\}^{-1}$, so that (B.4) approximates the form

$$p(y|n) \simeq \frac{(1 - r)(\nu_0 m_0)^{\nu_0/2} \Gamma((\nu_2/2) + 1 - r) \Gamma(\{(\nu_0 + \nu_1 - 1)/2\} + r)}{\pi^{nS/2} (cnS)^{\frac{1}{2}} \Gamma(\nu_0/2)(\nu_0 m_0 + \nu_1 m_1)^{(\nu_0 + \nu_1 - 1 + 2r)/2} (\nu_2 m_2)^{(\nu_2 + 2 - 2r)/2}}$$

$$\times I_w\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1 - 1}{2} + r\right). \tag{B.6}$$

In (3.4) and (3.5), the limiting improper case $r \to 1$ in the prior distribution of ϕ presented a problem in (A.4) when we tried to compute $p(\boldsymbol{x}|k) = p(y|n)p(y^*|n^*)$ (see Section 3.3, last paragraph). A similar problem occurs in (B.6) when $c \to \infty$, because the factor c^{-1} when $k > 1$ is much smaller than the corresponding factor $c^{-\frac{1}{2}}$ when $k = 1$. If we allow this to stand, the likelihood (B.6) for $k = 1$ will be infinitely greater than the corresponding likelihood for any value of $k > 1$. This difficulty can be resolved in one of two ways. Either we can keep c finite and do the numerical integration indicated by (B.5), or we can let $c \rightarrow \infty$ but put zero prior probability on the case $k = 1$. In this article we deal only with the latter option, for the sake of simplicity. Now k takes on the values $2, 3, \ldots, T$ with any set of prior probabilities, and $p(\boldsymbol{x}|k)$ always contains a factor of c^{-1}. This factor cancels out from numerator and denominator in the computation

$$p(k|\boldsymbol{x}) = \frac{p(k)p(\boldsymbol{x}|k)}{\Sigma_{k=2}^{T} p(k)p(\boldsymbol{x}|k)}.$$

In this situation we can also choose the diffuse prior on ϕ given by $p(\phi) \propto \phi^{-1}$, $0 < \phi \leq 1$, if we wish.

The formulae beginning with (A.5) for posterior and predictive means and variances hold for all versions of the model specifications of this section, except that the posterior moments of ϕ, given the break point, are given as follows for the case $c \rightarrow \infty$:

$$E(\phi^q|\boldsymbol{y}) = \frac{(1-w)^q B\left(\frac{\nu_2}{2} + q + 1 - r, \frac{\nu_0 + \nu_1 + 1}{2} - (q + 1 - r)\right) I_w\left(\frac{\nu_2}{2} + q + 1 - r, \frac{\nu_0 + \nu_1 + 1}{2} - (q + 1 - r)\right)}{w^q B\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1 + 1}{2} - (1 - r)\right) I_w\left(\frac{\nu_2}{2} + 1 - r, \frac{\nu_0 + \nu_1 + 1}{2} - (1 - r)\right)}$$

Note that in this section the between-site degrees of freedom ν_2 and the between-site mean square m_2 are different from Appendix A. Here we have $\nu_2 = S - 1$ and $\nu_2 m_2 = n\Sigma_s(\bar{y}_s - \bar{y}.)^2$.

REFERENCES

Box, G.E.P., and Tiao, G.C. (1968). Bayesian estimation of means for the random-effect model. *J. Amer. Statist. Assoc.*, 63, 174–181.

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, 70, 311–319.

Griffiths, William E., and Anderson, Jock R. (1982). Using time-series and cross-section data to estimate a production function with positive and negative marginal risks. *J. Amer. Statist. Assoc.*, 77, 529–540.

Jackson, Paul H.; Novick, Melvin R., and Thayer, Dorothy T. (1971). Estimating regressions in *m* groups. *British J. Math. Statist. Psych.*, 24, 129–153.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika*, 63, 69–75.

Lindley, D.V., and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34, 1–42.

Morris, Carl N. (1983). Parametric empirical Bayes inference: Theory and applications, with discussion. *J. Amer. Statist. Assoc.*, 78, 47–65.

Mosteller, Frederick, and Wallace, David L. (1964). *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Mass.

Rubin, Donald B. (1980). Using empirical Bayes techniques in the law school validity studies (with discussion). *J. Amer. Statist. Assoc.*, 75, 810–827.

Rubin, Donald B., and Stroud, T.W.F. (1977). Comparing high schools with respect to student performance in university, *J. Educ. Statist.*, 2, 139–155.

Rubin, Donald B., and Stroud, T.W.F. (1984). Bayesian forecasting in parallel time series, with application to university admissions. Preprint No. 1984-7, Dept. of Mathematics and Statistics, Queen's Univ. at Kingston.

Smith, A.F.M. (1980). Change-point problems: Approaches and applications. *Trabajos de Estadist.*, 31 (no. extra), 83–98.

Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.*, 42, 385–388.

Stroud, T.W.F. (1984). Bayesian shrinkage estimates for regression coefficients in *m* populations. *Comm. Statist. A—Theory Methods*, 13(17), 2085–2109.

Swamy, P.A.V.B., and Mehta, J.S. (1975). Bayesian and non-Bayesian analysis of switching regressions and of random coefficient regression models. *J. Amer. Statist. Assoc.*, 70, 593–602.

Swamy, P.A.V.B., and Mehta, J.S. (1977). Estimation of linear models with time and cross-sectionally varying coefficients. *J. Amer. Statist. Assoc.*, 72, 890–898.

Thisted, Ronald A., and Wecker, William E. (1981). Predicting a multitude of time series. *J. Amer. Statist. Assoc.*, 76, 516–523.

West, Mike; Harrison, P. Jeff, and Migon, Helio S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Assoc.*, 80, 73–97.

Received 17 September 1985
Revised 2 May 1986
Accepted 24 June 1986

Department of Mathematics and Statistics
Queen's University
Kingston, Ontario K7L 3N6

Department of Statistics
Harvard University
1 Oxford Street
Cambridge, MA 02138, U.S.A.